Capitolo 1

Matrici statistiche

1.1 Introduzione e notazioni preliminari

Un prerequisito indispensabile alle elaborazioni statistiche è la raccolta delle informazioni su un dato fenomeno.

Definizione 1.1 Si definisce *universo statistico* o *popolazione* l'insieme Ω di tutti gli elementi che si vogliono esaminare rispetto ad un dato fenomeno. I singoli elementi della popolazione $\omega \in \Omega$ si chiamano *unità statistiche*.

Prima di qualunque indagine risulta essenziale delimitare con precisione la popolazione che si vuole esaminare. Ciascuno degli individui della popolazione può essere considerato dal punto di vista di uno o più caratteri.

Definizione 1.2 Data una popolazione Ω, si chiama *carattere* X un'applicazione definita in Ω ed a valori in uno spazio $\mathcal{X}$, e scriveremo $X : \Omega \to \mathcal{X}$. Gli elementi di $\mathcal{X}$ si chiamano *modalità*.

Considerato un carattere $X : \Omega \to \mathcal{X}$, in generale possiamo distinguere due importanti casi:

1. $\mathcal{X}$ è un insieme finito di categorie e scriveremo $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$. Ad esempio: il formato di un file, tipo di scheda grafica, etc. Fra di esse distinguiamo caratteri qualitativi non ordinabili o di tipo *nominale* (come ad esempio forme, residenza) e caratteri ordinabili o di tipo *ordinale* (come ad esempio valutazioni di qualità o gradi di una gerarchia). In questo caso i caratteri vengono anche chiamati *variabili statistiche categoriali*.

2. $\mathcal{X}$ è un sottoinsieme di $\mathbb{R}$, cioè $\mathcal{X} \subseteq \mathbb{R}$. Ad esempio: dimensione di un file, quantità di memoria RAM, dimensioni monitor, velocità di comunicazione di un modem, etc. In questo caso bisogna ulteriormente distinguere il caso in cui $\mathcal{X}$ è un insieme al più numerabile, cioè $\mathcal{X} = \{x_1, x_2, \ldots, x_k, \ldots\}$ con $x_1 < x_2 < \cdots < x_k < \cdots$, dal caso in cui $\mathcal{X}$ coincide con un intervallo di $\mathbb{R}$, cioè del tipo $\mathcal{X} = [a, b]$, o addirittura con tutto $\mathbb{R}$, cioè $\mathcal{X} = \mathbb{R}$. Nel primo caso parliamo di *variabili statistiche discrete*, mentre nel secondo caso si parla di *variabili statistiche continue*.

Usualmente l'aggettivo "statistica" viene sottinteso e si parla di variabili categoriali, variabili discrete e variabili continue. Gli elementi di $\mathcal{X}$ vengono chiamati modalità della variabile (categoriale o quantitativa) X.

1.2 Matrici di dati

Sia $\Omega = \{\omega_1, \ldots, \omega_n\}$ una popolazione finita di n unità statistiche e si supponga di considerne lo studio attreverso p variabili statistiche $X_1, \ldots, X_p$ a valori rispettivamente in $\mathcal{X}_1, \ldots, \mathcal{X}_p$. Dopo aver completato tutte le fasi della rilevazione discusse in precedenza, viene costruita la *matrice dei dati*:

$$\begin{array}{c} \\ \omega_1 \\ \vdots \\ \omega_h \\ \vdots \\ \omega_n \end{array} \begin{array}{c} \begin{array}{ccccccc} X_1 & \cdots & X_i & \cdots & X_p \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{h1} & \cdots & x_{hi} & \cdots & x_{hp} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{ni} & \cdots & x_{np} \end{pmatrix} \end{array} .$$

in cui la colonna i-esima fornisce la distribuzione della variabile X_i fra le n unità statistiche rilevate; la riga j-esima fornisce le modalità rilevate dalle p variabili $X_1, \ldots, X_p$ nella j-esima unità statistica ω_j Matrici di dati di questo tipo vengono anche chiamate matrici *unità per variabili*.

Il file dati *alberghi*. La tabella seguente riporta i dati relativi ad alcuni alberghi nella provincia di Catanzaro rilevati nel 2001[1]. Si tratta di una matrice avente $n = 15$ righe (unità statistiche = alberghi) e $p = 5$ variabili:
X_1 = Comune di residenza (carattere qualitativo sconnesso);
X_2 = Ragione sociale (carattere qualitativo sconnesso);
X_3 = Categoria (carattere qualitativo ordinale);
X_4 = Posti letto (carattere quantitativo);
X_5 = Numero stanze (carattere quantitativo).

Si noti che nell'ultima riga, che riguarda il *Camping Residence Costa Blu*, non sono stati registrati né la categoria, né il numero di posti letti. né il numero di stanze. Si tratta di dati mancanti su cui torneremo in seguito.

La matrice di dati completa è contenuta nel file *alberghi.xls* insieme al tracciato record che contiene le informazioni sulle variabili rilevate e sulle fonti da cui provengono i dati, che in questo caso sono: annuario alberghi, elenchi Telecom, Pagine gialle Seat, Pagine utili.

[1]Fonte: CCIAA Catanzaro, Dati anno 2001, file: *alberghi.xls*; disponibile alla pagina: www.economia.unical.it/statistica/didattica/StatAziendale2/datiSA2.htm

unità	Comune	Ragione sociale	Categoria	P.letto	N. stanze
1	Sellia Marina	106	2	28	18
2	Cropani Marina	4 Lampioni	1	24	12
3	Platania	A Giurranda	2	40	20
4	Montepaone	A Lumera	3	68	35
5	Lamezia Terme	Aerhotel Phelipe	3	26	16
6	Montepaone	Alba Chiara Hotel Residence	3	80	40
7	Tiriolo	Autostello Chiarella	2	15	7
8	Stalettì	Baia Dell'Est	3	60	30
9	Catanzaro	Bellamena residence hotel	3	70	37
10	Badolato	Bell'orizzonte	2	80	40
11	Catanzaro	Belvedere	2	24	16
12	Catanzaro	Benny Hotel	3	152	76
13	Montepaone	Calaghena Hotel Villaggio	4	273	145
14	Decollatura	Caligiuri	3	18	10
15	Sellia Marina	Camping Residence Costa Blu			

Il file dati *german*. Un altro esempio di file dati concerne soggetti che hanno richiesto un credito presso un istituto bancario. Il file *german.xls*[2]. contiene 1000 rilevazioni per 20 diverse variabili; 700 individui si sono rivelati solventi buoni pagatori, mentre 300 si sono rivelati insolventi cattivi pagatori. Questa variabile è indicata con X_0: i primi costituiscono il gruppo "*good*" contrassegnato con 1, i secondi il gruppo "*bad*" contrassegnato con 2. I crediti in questione sono di piccolo importo (importo massimo circa pari a 16000 DM) e breve durata (durata massima 60 mesi). Le informazioni relative ai richiedenti credito sono descritte da 20 variabili (attributi), 7 continue e 13 categoriche.

Le varibili sono così definite:

- Variabile X_1: Saldo attuale del conto corrente, variabile categorica ordinale a 4 modalità (DM = marchi tedeschi):
  - A_{11}: $X_1 < 0DM$
  - A_{12}: $0 \leq X_1 < 200DM$
  - A_{13}: $X_1 \geq 200DM$
  - A_{14}: nessun conto corrente
- Variabile X_2: Durata del prestito in mesi, variabile quantitativa
- Variabile X_3: Storia del credito, variabile categorica nominale a 5 modalità

[2] fi le: *alberghi.xls*; disponibile alla pagina:
www.economia.unical.it/statistica/didattica/StatAziendale2/datiSA2.htm

  - A_{30}: nessun credito richiesto/tutti i crediti pagati in tempo
  - A_{31}: tutti i crediti richiesti a questa banca pagati in tempo
  - A_{32}: crediti attuali fin ora pagati in tempo
  - A_{33}: ritardi nei pagamenti in impegni passati
  - A_{34}: conto critico/altri crediti esistenti (non in questa banca)

- Variabile X_4: Scopo della richiesta di credito, variabile categorica nominale a 11 modalità

  - A_{40}: auto (nuova)
  - A_{41}: auto (usata)
  - A_{42}: mobilio/attrezzature
  - A_{43}: radio/televisione
  - A_{44}: elettrodomestici
  - A_{45}: riparazioni
  - A_{46}: istruzione
  - A_{47}: vacanza
  - A_{48}: riqualificazione
  - A_{49}: affari
  - A_{410}: altro

- Variabile X_5: Ammontare del credito richiesto, variabile quantitativa

- Variabile X_6: Tipologia conto di risparmio/obbligazioni, variabile categorica ordinale a 5 modalità

  - A_{61}: $X_6 < 100DM$
  - A_{62}: $100 \leq X_6 < 500DM$
  - A_{63}: $500 \leq X_6 < 1000DM$
  - A_{64}: $X_6 \geq 1000DM$
  - A_{65}: nessun conto di risparmio/sconosciuto

- Variabile X_7: Anni di assunzione dell'attuale impiego, variabile categorica ordinale a 5 modalità

  - A_{71}: disoccupato
  - A_{72}: $X_7 < 1$ anno
  - A_{73}: $1 \leq X_7 < 4$ anni
  - A_{74}: $4 \leq X_7 \leq 7$ anni
  - A_{75}: $X_7 \geq 7$ anni

- Variabile X_8: Stima della rata in percentuale all'entrata disponibile, variabile quantitativa

- Variabile X_9: Stato civile e sesso, variabile categorica nominale a 5 modalità

  – A_{91}: maschio: divorziato/separato
  – A_{92}: femmina: divorziata/separata/sposata
  – A_{93}: maschio: single
  – A_{94}: maschio: sposato/vedovo
  – A_{95}: femmina: single

- Variabile X_{10}: Altri debitori/garanti, variabile categorica nominale a 3 modalità

  – A_{101}: nessuno
  – A_{102}: co-richiedente
  – A_{103}: garante

- Variabile X_{11}: Anni dell'attuale residenza, variabile quantitativa

- Variabile X_{12}: Proprietà, variabile categorica nominale, 4 modalità

  – A_{121}: beni immoboli
  – A_{122}: se non è A_{121}: società di credito edilizio/assicurazione sulla vita
  – A_{123}: se non è $A_{121} o A_{122}$: auto o altro, non presente nella variabile 6
  – A_{124}: sconosciute/nessuna proprietà

- Variabile X_{13}: Età, variabile quantitativa

- Variabile X_{14}: Altri schemi di pagamento, variabile categorica nominale a 3 modalità

  – A_{141}: banca
  – A_{142}: negozi
  – A_{143}: nessuno

- Variabile X_{15}: Casa di abitazione, variabile categorica a tre modalità

  – A_{151}: affitto
  – A_{152}: di proprità
  – A_{153}: ospite

- Variabile X_{16}: Numero di crediti attualmente in questa banca, variabile quantitativa

- Variabile X_{17}: Lavoro, variabile categorica nominale a 4 modalità

  – A_{171}: disoccupato/non-specializzato, non-residente
  – A_{172}: non specializzato residente

  - A_{173}: impiegato specializzato/funzionario
  - A_{174}: dirigente/autonomo/impiego altamente qualificato/ufficiale

- Variabile X_{18}: Numero di persone a carico, variabile quantitativa

- Variabile X_{19}: Telefono, variabile categorica nominale a 2 modalità

  - A_{191}: no
  - A_{192}: si

- Variabile X_{20}: Lavoratore straniero, variabile categorica nominale a 2 modalità

  - A_{201}: si
  - A_{202}: no.

Si noti che le variabili possono essere raggruppate in tre categorie, relativamente a informazioni su:

1. dati personali e tipo di lavoro (var.: $X_7, X_9, X_{11}, X_{13}, X_{17}, X_{18}, X_{19}, X_{20}$)
2. stato economico del soggetto (var.: $X_1, X_3, X_6, X_{10}, X_{12}, X_{14}, X_{15}, X_{16}$)
3. tipo di credito (var.: X_2, X_4, X_5, X_8)

Caso generale. In generale, una matrice di dati $n \times p$ verrà denotata col simbolo $\underset{\sim}{\mathbf{X}}$ o $\underset{n\times p}{\mathbf{X}}$, l'elemento della i-esima riga e j-esima colonna è x_{ij} e denota il valore osservato per la j-esima variabile nella i-esima unità statistica. Nel caso precedente si tratta di una matrice di dati 15×5.

In base a quanto visto in precedenza, le righe di $\underset{\sim}{\mathbf{X}}$ possono essere scritte come $\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'$[3] e le colonne di $\underset{\sim}{\mathbf{X}}$ verranno scritte con l'indice fra parentesi, cioè $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(p)}$. Riassumendo, possiamo scrivere:

$$\underset{\sim}{\mathbf{X}} := \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(p)}) = (x_{hi})$$

dove

$$\mathbf{x}_h := \begin{pmatrix} x_{h1} \\ x_{h2} \\ \vdots \\ x_{hp} \end{pmatrix} \quad \text{e} \quad \mathbf{x}_{(i)} := \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix}$$

per $h = 1, 2, \ldots, n$ e $i = 1, 2, \ldots, p$.

[3] In genere con $\mathbf{x}$ si intende un vettore colonna e con $\mathbf{x}'$ o $\mathbf{X}'$ denota rispettivamente vettore trasposto o matrice trasposta.

Si noti che $\mathbf{x}_i$ è il vettore p-dimensionale che denota le p osservazioni sulla i-esima unità statistica, mentre $\mathbf{x}_{(i)}$ è il vettore n dimensionale che denota le n osservazioni sulla i-esima . Ad esempio, per i dati sugli alberghi:

$$\mathbf{x}_7 = (\text{Tiriolo}, \text{Autostello Chiarella}, 2, 15, 7)'$$
$$\mathbf{x}_{12} = (\text{Catanzaro}, \text{Benny Hotel}, 3, 152, 76)'$$

e

$$\mathbf{x}_{(1)} = (\text{Sellia Marina, Cropani Marina, Platania, Montepaone, Lamezia Terme, Montepaone, Tiriolo, Stalettì, Catanzaro, Badolato, Catanzaro, Catanzaro, Montepaone, Decollatura, Sellia Marina})'$$
$$\mathbf{x}_{(3)} = (2, 1, 2, 3, 3, 3, 2, 3, 3, 2, 2, 3, 4, 3, \text{NA})'$$

dove NA denota "dato mancante" (not avalaible).

Nell'analisi multivariata, le righe $\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_n$ formano usualmente un campione casuale, al contrario ciò non si verifica per le colonne $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(p)}$. Chiaramente quando il numero n di unità rilevate ed il numero p di caratteri da analizzare è moderatamente grande, il numero totale di informazioni np può risultare molto grande, per cui risulta essenziale sintetizzare in maniera opportuna i dati rilevati.

Infine rileviamo che le matrici di dati possono essere studiate secondo due diversi punti di vista: l'analisi delle colonne conduce allo studio della relazione fra le variabili; l'analisi delle righe conduce allo studio delle relazioni fra unità statistiche differenti.

1.2.1 Matrici di frequenza (tabelle a doppia entrata)

Consideriamo il caso in cui la popolazione Ω viene studiata in base a due caratteri X e Y aventi rispettivamente h e k modalità, cioè $x_1, \ldots, x_k$ e $y_1, \ldots, y_h$. I dati possono essere riassunti in una *matrice di frequenza* o *tabella a doppia entrata* in cui a ciascuna coppia di modalità (x_i, y_j) di (X, Y) si fa corrispondere la sua frequenza assoluta n_{ij} $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, h$, cioè il numero di volte in cui la coppia di modalità si è presentata:

	y_1	$\ldots$	y_j	$\ldots$	y_h	
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1h}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{ih}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	n_{k1}	$\ldots$	n_{kj}	$\ldots$	n_{kh}	n_{k0}
	n_{01}	$\ldots$	n_{0j}	$\ldots$	n_{0h}	N

dove:

$$n_{i0} = \sum_{j=1}^{h} n_{ij} , \qquad n_{0j} = \sum_{i=1}^{k} n_{ij} \qquad \text{e } N = \sum_{i=1}^{k} \sum_{j=1}^{h} n_{ij}$$

Le frequenze N_{i0}, $i = 1, 2, \ldots, k$ vengono chiamate *frequenze marginali assolute* di X; le frequenze N_{0j}, $j = 1, 2, \ldots, h$ vengono chiamate *frequenze marginali assolute* di Y ; N è ovviamente il numero di elementi della popolazione.

Ad esempio, la seguente matrice di frequenza fornisce il numero di posti di lavoro creati dalle imprese italiane nel 1998 per ramo di attività produttiva e zona geografica[4]:

Ramo	*Nord*	*Centro*	*Sud e Isole*	TOTALE
Energia, gas, acqua	198	45	79	322
Industrie estrattive e chimiche	4.248	1.417	1.893	7.558
Industrie manifatturiere dei metalli	27.683	5.688	4.991	38.362
Altre industrie manifatturiere	23.063	9.898	8.958	41.919
Industrie per l'edilizia	19.885	8.321	16.101	44.307
Commercio, pubblici esercizi e alberghi	41.208	16.088	18.289	75.585
Trasporto e comunicazioni	4.764	1.748	2.635	9.147
Credito e assicurazione	13.265	4.423	3.836	21.524
Servizi pubblici e privati	11.238	4.407	4.749	20.394
TOTALE	134.314	47.628	56.782	238.724

Ovviamente possono essere introdotte anche in questo caso le frequenze relative percentuali:

$$f_{ij} \;=\; \frac{n_{ij}}{N} \qquad\qquad i = 1, 2, \ldots, k, \qquad j = 1, 2, \ldots, h \tag{1.1}$$

e quindi riscrivere la tabella a doppia entrata in funzione di tali frequenze:

	y_1	$\ldots$	y_j	$\ldots$	y_h	
x_1	f_{11}	$\ldots$	f_{1j}	$\ldots$	f_{1h}	f_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\ldots$	f_{ij}	$\ldots$	f_{ih}	f_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	f_{k1}	$\ldots$	f_{kj}	$\ldots$	f_{kh}	f_{k0}
	f_{01}	$\ldots$	f_{0j}	$\ldots$	f_{0h}	100

dove:

$$f_{i0} \;=\; \sum_{j=1}^{h} f_{ij} \qquad \text{e} \;\; f_{0j} \;=\; \sum_{i=1}^{k} f_{ij} \;.$$

Le frequenze f_{i0}, $i = 1, 2, \ldots, k$ vengono chiamate *frequenze marginali relative* di X; le frequenze f_{0j}, $j = 1, 2, \ldots, h$ vengono chiamate *frequenze marginali relative* di Y.

Considerando le frequenze relative percentuali, la precedente tabella si scrive:

[4]Fonte: Banca dati INPS, *www.inps.it.*

Ramo	*Nord*	*Centro*	*Sud e Isole*	TOTALE
Energia, gas, acqua	0,08	0,02	0,03	0,13
Industrie estrattive e chimiche	1,78	0,59	0,79	3,17
Industrie manifatturiere dei metalli	11,60	2,38	2,09	16,07
Altre industrie manifatturiere	9,66	4,15	3,75	17,56
Industrie per l'edilizia	8,33	3,49	6,74	18,56
Commercio, pubblici esercizi e alberghi	17,26	6,74	7,66	31,66
Trasporto e comunicazioni	2,00	0,73	1,10	3,83
Credito e assicurazione	5,56	1,85	1,61	9,02
Servizi pubblici e privati	4,71	1,85	1,99	8,54
TOTALE	56,26	19,95	23,79	100,00

1.2.2 Matrice di origine-destinazione

La matrice di origine-destinazione è una matrice di frequenze in cui i valori n_{ij} sono le frequenze degli spostamenti dalla posizione sulla riga i-esima alla posizione sulla colonna j-esima.

Ad esempio, la seguente tabella fornisce il numero di arrivi degli italiani negli esercizi ricettivi per le regioni di provenienza e destinazione nell'area sud e isole per l'anno 2000[5]

Regione di	*Regione di arrivo*					
provenienza	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
Campania	775,337	116,371	41,356	120,367	136,728	39,322
Puglia	134,421	208,380	60,197	88,092	88,739	16,796
Basilicata	48,034	17,188	26,866	13,677	12,835	2,810
Calabria	69,534	50,355	13,401	109,307	87,739	5,696
Sicilia	90,594	113,284	15,286	116,679	870,498	22,023
Sardegna	21,449	14,077	1,367	3,128	21,860	221,201

La matrice completa è riportata nel file *turisti.xls*[6].

Altri esempi di matrici origine-destinazione riguardano: il movimento di pazienti che vengono curati in strutture sanitarie della stessa regione di residenza oppure diversa; clienti di telefonia mobile che continuano ad usufruire del servizio di una stesso gestore di rete oppure lo cambiano.

[5]Fonte: ISTAT, *Statistiche del Turismo*, anno 2000, tav. 2.46.

[6]disponibile alla pagina: www.economia.unical.it/statistica/didattica/StatAziendale2/datiSA2.htm

1.3 Stastistiche descrittive per matrici di dati quantitativi

Quando la matrice di dati contiene unicamente variabili quantitative, è possibile effettuare opportune sintesi, che sono essenzialmente generalizzazioni delle statistiche nel caso univariato e bivariato.

Consideriamo la seguente matrice di dati che concerne lo studio dei vini di di Bordeaux in base alla loro qualità. La tabella riporta 34 osservazioni climatiche dei mesi da aprile a settembre nelle annate dal 1924 al 1957, dove:

X_1 : somma delle temperature medie giornaliere in gradi;
X_2 : durata del sole in ore;
X_3 : numero di giorni di grande caldo;
X_4 : quantità di pioggia in mm.

La tabella qui di seguito riporta solo i primi dieci anni, i dati completi si trovano nel file *bordeaux.xls*[7].

Annata	X_1	X_2	X_3	X_4	Qualità
24	3064	1201	10	361	+
25	3000	1053	11	338	−
26	3155	1133	19	393	+
27	3085	970	4	467	−−
28	3245	1258	36	294	++
29	3267	1386	35	225	+++
30	3080	966	13	417	−−
31	2974	1185	12	488	−−
32	3038	1103	14	677	− − −
33	3318	1310	29	427	+

L'ultima variabile fornisce una valuazione di qualità, secondo 7 categorie da $+++$ a $---$ ed è stata inserita successivamente in base al giudizio espresso da alcuni *sommeliers*.

Il vettore delle medie e la matrice di covarianza Un'ovvia estensione delle nozioni univariate di media (aritmetica) e varianza conduce alle definizioni seguenti. Se si considera la generica variabile X_i, $i = 1, 2, \ldots, p$, essa avrà media e varianza data rispettivamente da:

$$\bar{x}_i := \frac{1}{n}\sum_{h=1}^{n} x_{hi} \tag{1.2}$$

$$s_{ii} := \frac{1}{n}\sum_{h=1}^{n}(x_{hi} - \bar{x}_i)^2 \;=\; \frac{1}{n}\sum_{h=1}^{n} x_{hi}^2 - \bar{x}_i^2 \;=\; s_i^2 \tag{1.3}$$

per $i = 1, 2, \ldots, p$.

[7] disponibile alla pagina: www.economia.unical.it/statistica/didattica/StatAziendale2/datiSA2.htm

Assegnata una coppia di variabili X_i, X_j, con $i, j = 1, 2, \ldots, p$, la covarianza fra X_i e X_j è definita da:

$$s_{ij} := \frac{1}{n}\sum_{h=1}^{n}(x_{hi} - \bar{x}_i)(x_{hj} - \bar{x}_j) \ . \tag{1.4}$$

I valori delle medie di tutte le p variabili $X_1, X_2, \ldots, X_p$ possono essere descritti sinteticamente mediante un vettore p-dimensionale, chiamato *vettore delle medie campionarie*, avente per coordinate le medie delle singole variabili:

$$\bar{\mathbf{x}} := \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \ . \tag{1.5}$$

Per esempio, con riferimento all'intera matrice di dati per il caso precedente, si ha:

$$\mathbf{x} = (3158.76, 1248.71, 18.82, 360.44)'$$

Analogamente la covarianza fra tutte le possibili coppie di variabili X_i, X_j, con $i, j = 1, 2, \ldots, p$, può essere sinteticamente data mediante una matrice quadrata **S** di ordine $p \times p$ chiamata *matrice di varianze e covarianze*:

$$\mathbf{S} := (s_{ij}) \ ,$$

dove, in particolare, $s_{ii} = s_i^2$ è la varianza della i-esima variabile. Essendo $s_{ij} = s_{ji}$, segue che **S** è una matrice simmetrica. Nel caso dei vini di Bordeaux si ha:

$$\mathbf{S} = \begin{pmatrix} 19693.24 & 12549.40 & 1192.87 & 8108.54 \\ 12549.40 & 15655.91 & 797.54 & -5352.28 \\ 1192.87 & 797.54 & 97.38 & -356.45 \\ -5203.78 & -5352.28 & -356.45 & 8108.54 \end{pmatrix} \ .$$

Le statistiche precedenti possono essere espresse in notazione matriciale come segue. In base a (1.2) e (1.5) si ha:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{h=1}^{n}\mathbf{x}_h = \frac{1}{n}\underset{\sim}{\mathbf{X}}'\mathbf{1} \tag{1.6}$$

dove $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^n$ è il vettore unitario n dimensionale. Inoltre dalla (1.4) si ha:

$$s_{ij} = \frac{1}{n}\sum_{h=1}^{n}x_{hi}x_{hj} - \bar{x}_i\bar{x}_j \ . \tag{1.7}$$

Notiamo che il termine $(x_{hi} - \bar{x}_i)(x_{hj} - \bar{x}_j)$, per $h = 1, \ldots, n$, può essere visto come l'elemento ij della matrice $(\mathbf{x}_h - \bar{\mathbf{x}})(\mathbf{x}_h - \bar{\mathbf{x}})'$; analogamente il termine $x_{hi}x_{hj}$ può essere visto come l'elemento ij della matrice $\mathbf{x}_h\mathbf{x}_h'$. Si ha infatti, ad esempio:

$$\mathbf{x}_h\mathbf{x}_h' = \begin{pmatrix} x_{h1} \\ x_{h2} \\ \vdots \\ x_{hp} \end{pmatrix}(x_{h1}, x_{h2}, \cdots, x_{hp}) = \begin{pmatrix} x_{h1}^2 & x_{h1}x_{h2} & \cdots & x_{h1}x_{hp} \\ x_{h2}x_{h1} & x_{h2}^2 & \cdots & x_{h2}x_{hp} \\ \cdots & \cdots & \cdots & \cdots \\ x_{hp}x_{h1} & x_{hp}x_{h2} & \cdots & x_{hp}^2 \end{pmatrix} = (x_{hi}x_{hj}) \ .$$

Allora, per le (1.4) e (1.7), possiamo scrivere:

$$\underset{p\times p}{\mathbf{S}} := (s_{ij}) = \frac{1}{n}\sum_{h=1}^{n}(\mathbf{x}_h - \bar{\mathbf{x}})(\mathbf{x}_h - \bar{\mathbf{x}})' = \frac{1}{n}\sum_{h=1}^{n}\mathbf{x}_h\mathbf{x}_h' - \bar{\mathbf{x}}\bar{\mathbf{x}}' \ . \tag{1.8}$$

Tenendo conto che $\sum_{h=1}^{n}\mathbf{x}_h\mathbf{x}_h' = \underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}$, sostituendo, in base alla (1.6), segue:

$$\mathbf{S} = \frac{1}{n}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}} - \bar{\mathbf{x}}\bar{\mathbf{x}}' = \frac{1}{n}\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}} - \frac{1}{n}\underset{\sim}{\mathbf{X}}'\mathbf{1}\mathbf{1}'\underset{\sim}{\mathbf{X}}\right) \ . \tag{1.9}$$

Se introduciamo la *matrice centrante* **H**:

$$\underset{n\times n}{\mathbf{H}} := \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' = \mathbf{I} - \frac{1}{n}\mathbf{J} \tag{1.10}$$

dove **I** è la matrice identità e $\mathbf{J} = \mathbf{1}\mathbf{1}'$, si ottiene:

$$\mathbf{S} = \frac{1}{n}\underset{\sim}{\mathbf{X}}'\mathbf{H}\underset{\sim}{\mathbf{X}} \ , \tag{1.11}$$

che costituisce una conveniente rappresentazione della matrice di varianze e covarianze.

La matrice **H** è una matrice simmetrica e idempotente. Infatti dalla (1.10) si vede immediatamente che è combinazione lineare di matrici simmetriche; inoltre:

$$\begin{aligned}\mathbf{H}^2 &= (\mathbf{I} - \frac{1}{n}\mathbf{J})(\mathbf{I} - \frac{1}{n}\mathbf{J}) = \mathbf{I} - \frac{2}{n}\mathbf{J} + \frac{1}{n^2}\mathbf{J}^2 = \mathbf{I} - \frac{2}{n}\mathbf{J} + \frac{1}{n^2}n\mathbf{J} = \mathbf{I} - \frac{2}{n}\mathbf{J} + \frac{1}{n}\mathbf{J} \\ &= \mathbf{I} - \frac{1}{n}\mathbf{J} = \mathbf{H} \, .\end{aligned}$$

Ne segue che per ogni vettore $\mathbf{a} \in \mathbb{R}^p$ si ha:

$$\mathbf{a}'\mathbf{S}\mathbf{a} = \frac{1}{n}\mathbf{a}'\underset{\sim}{\mathbf{X}}'\mathbf{H}'\mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{a} = \frac{1}{n}\mathbf{y}\mathbf{y}' \ \geq \ 0$$

dove $\mathbf{y} = \mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{a}$. Pertanto la matrice di varianze e covarianze **S** è semidefinita positiva ($\mathbf{S} \geq 0$). Per dati continui usualmente la matrice **S** risulta essere definita positiva se risulta $n \geq p + 1$.

La matrice di correlazione. Il coefficiente di correlazione lineare semplice fra due variabili X_i e X_j è definito da

$$r_{ij} := \frac{s_{ij}}{s_i s_j} \ . \tag{1.12}$$

dove, come ben noto, si ha $-1 \leq r_{ij} \leq 1$, e $r_{ii} = 1$.

Analogamente a quanto visto in precedenza, si può quindi introdurre la *matrice di correlazione* **R**:

$$\mathbf{R} := (r_{ij}) \ . \tag{1.13}$$

La matrice di correlazione, in base alla (1.12), è quindi simmetrica. Nel caso dei vini di Bordeaux si ha:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.7147 & 0.8614 & -0.4118 \\ 0.7147 & 1.0000 & 0.6459 & -0.4750 \\ 0.8614 & 0.6459 & 1.0000 & -0.4011 \\ -0.4118 & -0.4750 & -0.4011 & 1.0000 \end{pmatrix} \ . \tag{1.14}$$

Posto $\mathbf{D}$ =diag(s_i), dove $s_i = \sqrt{s_{ii}}$, in funzione della matrice di covarianza si ha la seguente relazione che generalizza la definizione di coefficiente di correlazione lineare al caso matriciale:

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1} \ , \tag{1.15}$$

e pertanto $\mathbf{R}$ è semidefinita positiva (per un noto risultato dell'algebra lineare)[8]

In particolare se $\mathbf{R} = \mathbf{I}$, diremo che le variabili sono non correlate.

La matrice di correlazione parziale. E' anche opportuno considerare la matrice delle correlazioni parziali, i cui elementi sono i coefficienti di correlazione tra due variabili a parità di tutte le altre, [Riz85].

Il generico elemento (i, j) di tale matrice viene indicato con $r_{ij,p}$, dove con p indichiamo l'insieme di tutte le variabili eccetto X_i e X_j. L'espressione generale di $r_{ij,p}$ è:

$$r_{ij,p} := \frac{-R_{ij}}{\sqrt{R_{ii}R_{jj}}} \tag{1.16}$$

dove R_{ij} è il complemento algebrico o aggiunto[9] di r_{ij} nella matrice di correlazione $\mathbf{R}$.

Per comprendere il significato del coefficiente di correlazione parziale (1.16) consideriamo il caso $p = 3$ e le variabili X_1, X_2 e X_3. Preliminarmente, ricordiamo che il coefficiente di regressione lineare fra due variabili X_1 e X_2 può essere visto come la media geometrica dei due coefficienti di regressione b_{12} e b_{21} rispettivamente della retta di regressione di X_1 su X_2 e di quella di X_2 su X_1:

$$r = \sqrt{b_{12} \cdot b_{21}}$$

Supponiamo di aver determinato Consideriamo l'equazione di regressione di X_1 rispetto a X_2 e X_3 e quella di X_2 rispetto a X_1 e X_3:

$$\begin{aligned} X_1 &= f(x_2, x_3) + \epsilon_1 = b_{12}x_2 + b_{13}x_3 + b_{10} + \epsilon_1 \\ X_2 &= f(x_1, x_3) + \epsilon_2 = b_{21}x_1 + b_{23}x_3 + b_{20} + \epsilon_2 \ , \end{aligned}$$

[8] **Teorema.** Sia $\mathbf{A} \geq 0$ una matrice $p \times p$. Allora per ogni matrice $\mathbf{C}$ $p \times n$, la matrice $\mathbf{C}'\mathbf{A}\mathbf{C}$ `e semidefi nita positiva. In particolare, se $\mathbf{A} > 0$ e $\mathbf{C}$ `e una matrice non singolare (e quindi $p = n$), allora $\mathbf{C}'\mathbf{A}\mathbf{C} > 0$.

[9] **Defi nizione.** Sia $\mathbf{A} = (a_{ij})$ una matrice quadrata di ordine $n \times n$. Si defi nisce *minore* dell'elemento a_{ij} il determinante della matrice ottenuta da $\mathbf{A}$ sopprimendo la i-esima riga e la j-esima colonna di $\mathbf{A}$. Si defi nisce *complemento algebrico* o *aggiunto* dell'elemento a_{ij}, e si denota col simbolo A_{ij}, il prodotto del minore di a_{ij} per $(-1)^{i+j}$.

Ad esempio, assegnata la matrice $\mathbf{A}(3 \times 3)$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

i minori degli elementi a_{11} e a_{12} sono dati rispettivamente da:

$$\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} \ , \qquad \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix}$$

ed i complementi algebrici A_{11} e A_{12} sono dati rispettivamente da:

$$A_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \ , \qquad A_{12} = -\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$$

dove b_{12} fornisce la variazione media di X_1 rispetto ad una variazione unitaria di x_2 per x_3 fissato, e b_{21} fornisce la variazione media di x_2 per un aumento unitario di x_1, con x_3 costante. Infine ϵ_1, ϵ_2 sono variabili aleatorie con media zero e varianza finita e costante.

Si può quindi misurare la concordanza o la discordanza fra X_1 e X_2, per X_3 costante, con la media geometrica dei due coefficienti di regressione parziale $b_{12,3}$ e $b_{21,3}$ che viene chiamata *coefficiente di correlazione parziale fra X_1 e X_2 con X_3 costante*. I valori $b_{12,3}$ e $b_{21,3}$ coincidono numericamente rispettivamente con b_{12} e b_{21} ma hanno significato statistico un po' diverso. Infatti, facendo assumere a X_3 valori diversi, si ottengono altrettante rette di regressione di X_1 su X_2 (ed analogamente di X_2 su X_1) che differiscono per l'ordinata all'origine (termine noto) ma che hanno sempre lo stesso valore di $b_{12,3}$ (risp. $b_{21,3}$), cioè lo stesso coefficiente angolare e pertanto sono parallele. Si intende per coefficiente di regressione parziale di X_1 su X_2 (risp. di X_2 su X_1) tenendo costante X_3 proprio questo valore comune.

Si ha pertanto:

$$r_{12,3} = \sqrt{b_{12,3} \cdot b_{21,3}} ,$$

dove $b_{12,3}$ è il coefficiente di regressione parziale di X_1 rispetto a X_2, fissato X_3 e $b_{21,3}$ è il coefficiente di regressione parziale di X_2 rispetto a X_1, fissato X_3. Tenendo conto delle relazioni fra coefficiente di regressione e coefficienti di correlazione nella regressione multipla, che sono dati rispettivamente da:

$$\begin{aligned} b_{12,3} &= \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\frac{\sigma_1}{\sigma_2} \\ b_{21,3} &= \frac{r_{21} - r_{13}r_{23}}{1 - r_{13}^2}\frac{\sigma_2}{\sigma_1} \end{aligned}$$

da cui la media geometrica è data da (si tenga conto della simmetria della matrice R, per cui $r_{12} = r_{21}$):

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} . \tag{1.17}$$

In termini matriciali, dalla (1.16) si ha:

$$r_{12,3} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}} = \frac{-(-)\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\sqrt{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}\begin{vmatrix} 1 & r_{13} \\ r_{13} & 1 \end{vmatrix}}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{23}^2)(1 - r_{13}^2)}}$$

che è proprio la (1.17).

Ritorniamo alla matrice dei dati dei vini di Bordeaux e calcoliamo $R_{12,34}$. Dalla (1.14)

si ha preliminarmente:

$$R_{12} = (1+2)^{-1} \begin{vmatrix} r_{21} & r_{23} & r_{24} \\ r_{31} & r_{33} & r_{34} \\ r_{41} & r_{43} & r_{44} \end{vmatrix} = - \begin{vmatrix} 0.7147 & 0.6459 & -0.4750 \\ 0.8614 & 1.0000 & -0.4011 \\ -0.4118 & -0.4011 & 1.0000 \end{vmatrix} = 0.1185$$

$$R_{11} = (1+1)^{-1} \begin{vmatrix} r_{22} & r_{23} & r_{24} \\ r_{32} & r_{33} & r_{34} \\ r_{42} & r_{43} & r_{44} \end{vmatrix} = \begin{vmatrix} 1.0000 & 0.6459 & -0.4750 \\ 0.6459 & 1.0000 & -0.4011 \\ -0.4750 & -0.4011 & 1.0000 \end{vmatrix} = 0.4424$$

$$R_{22} = (2+2)^{-1} \begin{vmatrix} r_{11} & r_{13} & r_{14} \\ r_{31} & r_{33} & r_{34} \\ r_{41} & r_{43} & r_{44} \end{vmatrix} = \begin{vmatrix} 1.0000 & 0.8614 & -0.4118 \\ 0.8614 & 1.0000 & -0.4011 \\ -0.4118 & -0.4011 & 1.0000 \end{vmatrix} = 0.2121$$

e quindi:

$$r_{12,34} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}} = \frac{0.1185}{\sqrt{0.4424 \cdot 0.2121}} = 0.3870$$

Effettuando tutti i calcoli, si ottiene la seguente matrice di correlazione parziale:

$$\mathbf{R}^* = \begin{pmatrix} 1.0000 & 0.3870 & 0.7447 & -0.0188 \\ 0.3870 & 1.0000 & 0.0595 & -0.2774 \\ 0.7447 & 0.0595 & 1.0000 & -0.0797 \\ -0.0188 & -0.2774 & -0.0797 & 1.0000 \end{pmatrix} . \tag{1.18}$$

Dal confronto con la (1.14) si può notare l'influenza delle altre variabili nella correlazione bivariata.

1.3.1 Misure di dispersione multivariata.

Come già osservato, la matrice di covarianza **S** è una generalizzazione della nozione di varianza introdotta nel caso univariato. Spesso comunque, anche nel caso multivariato, è conveniente avere un singolo numero che esprima una misura della dispersione dei dati.

Due misure che usualmente vengono considerate a tal fine sono

1. la *variazione totale* VAR_T: la traccia della matrice di covarianze tr**S**;

2. la *varianza generalizzata* VAR_G: il determinante della matrice di covarianza $|\mathbf{S}|$.

Per entrambe le misure, grandi valori indicano una grande variabilità intorno al punto medio $\bar{\mathbf{x}}$ e, al contrario, piccoli valori indicano che i dati sono abbastanza concentrati intorno al punto medio $\bar{\mathbf{x}}$. Ciascuna di tali misure, inoltre, riflette aspetti differenti della variabilità dei dati. e la varianza generalizzata svolge un ruolo importante nella stima di massima verosimiglianza.

La variazione totale. La traccia della matrice di covarianze corrisponde alla somma delle varianze della singole variabili $X_1, \ldots, X_p$:

$$\text{VAR}_T := \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \sigma_i^2 \ .$$

La varianza totale presenta problemi interpretativi se le variabili hanno diversa unità di misura ed inoltre ha il difetto di non tenere in conto della correlazione eventualmente presente fra le variabili.

La variazione totale è un concetto utile nell'analisi in componenti principali

La varianza generalizzata di Wilks. La varianza generalizzata è data dal determinante della matrice di covarianze:

$$|\mathbf{S}| := \text{det}\mathbf{S}$$

ed è una misura di variabilità che tiene conto della correlazione fra le variabili. In particolare la varianza generalizzata risulta uguale a zero se il rango[10] della matrice di covarianza è minore di p e tale caso si manifesta quando almeno una variabile assume sempre il medesimo valore nelle n unità statistiche (cioè risulta costante), oppure almeno una variabile è perfettamente correlata con un'altra, oppure se se una variabile è combinazione lineare di altre variabili. In queste circostanze, la variabilità non è a p dimensioni ma a $(p-1)$, $(p-2)$, ... dimensioni.

Il valore massimo che può assumere la varianza generalizzata è uguale al prodotto delle varianze delle p variabili, per cui si può definire la seguente misura relativa di variabilità multidimensionale:

$$V_R := \frac{|\mathbf{S}|}{\prod_{i=1}^{p} \sigma_i^2} \ .$$

Infine si può dimostrare che V_R è uguale al determinante della matrice di correlazione:

$$V_R |\mathbf{R}| \ .$$

Ovviamente V_R risulta uguale a zero nelle medesime circostanze in cui vale zero la varianza generalizzata ed assume valore massimo unitario quando le variabili sono fra loro incorrelate a due a due. Da ciò si deduce che, per valori fissi delle varianze delle p variabili, la varianza generalizzata aumenta quando diminuisce la correlazione fra le variabili.

[10] Sia **A** una matrice di ordine $n \times p$. Si defi nisce *rango* di **A**, denotato con $r(\mathbf{A})$, il massimo numero di righe (colonne) linearmente indipendenti di **A**.

Alcune propriet`a del rango di una matrice sono qui di seguito riassunte: Sia **A** una matrice di ordine $n \times p$. Allora si ha:

1. $0 \leq r(\mathbf{A}) \leq \min(n, p)$;
2. $r(\mathbf{A}) = r(\mathbf{A}')$;
3. $r(\mathbf{A}'\mathbf{A}) = r(\mathbf{A}\mathbf{A}') = r(\mathbf{A})$;
4. $r(\mathbf{A} + \mathbf{B}) \leq r(\mathbf{A}) + r(\mathbf{B})$ per ogni matrice $\mathbf{B}(n \times p)$;
5. $r(\mathbf{AB}) \leq \min\{r(\mathbf{A}), r(\mathbf{B})\}$ per ogni matrice $\mathbf{B}(p \times q)$;
6. $r(\mathbf{BAC}) = r(\mathbf{A})$ per tutte le matrici non singolari $\mathbf{B}(n \times n$ e $\mathbf{C}(p \times p$;
7. se $n = p$ allora $r(\mathbf{A}) = n$ se e solo se **A** `e non singolare.

1.4 Codifica dei dati

In molti casi, prima di applicare tecniche di analisi dei dati si ricorre ad una ricodifica dei dati rilevati per una o più variabili.

Consideriamo ad esempio il dataset *german*. La variabile X_4 (Saldo attuale del conto corrente) è una variabile categoriale ordinale a 4 modalità che possono essere codificate come: 1, 2, 3, 0 (se "sconosciuto", quest'ultima modalità può essere intesa come dato mancante). Si noti che la variabile originaria è quantitativa ed è stata ricodificata come categoriale ordinale.

La variabile X_4 (scopo del prestito) è una variabile categoriale nominale a 11 modalità che possono essere codificate come: $0, 1, 2, \ldots, 10$. Ovviamente il trattamento delle variabili dipende dalla loro natura e non dalla codifica. Ad esempio non ha senso fare la media di X_4, anche se viene rappresentata in forma numerica.

Un importante esempio di codifica è il seguente.

La codifica disgiuntiva completa. Consideriamo un carattere qualitativo che presenti r modalità (usualmente è $r \leq n$, ma solitamente $r \ll n$). La matrice di dati può prevedere che la corrispondente colonna fornisca i valori codificati di tali modalità. In alternativa le osservazioni possono essere codificate in maniera disgiuntiva sostituendo alla colonna in esame r colonne relative a *variabili indicatrici* X_v, $v = 1, \ldots, r$ ciascuna delle quali corrisponde ad una modalità di X e tali che:

$$x_n i \quad = \quad \begin{cases} 1 & \text{se la modalità } i \text{ è presente nell'}h\text{-esima unità} \\ 0 & \text{se la modalità } i \text{ è assente nell'}h\text{-esima unità.} \end{cases}$$

Consideriamo ad esempio i risultati di un'indagine sul diploma di Scuola Media Superiore conseguito nel 1998 da alcuni studenti iscrittisi successivamente coltà di Economia:

Unità	*Classico*	*Scientifico*	*Commerciale*	*Altre*
1	0	0	1	0
2	0	1	0	0
3	0	0	1	0
4	1	0	0	0
5	0	0	1	0
6	0	0	1	0
7	0	0	1	0
8	0	0	1	0
9	0	1	0	0
10	0	0	1	0
11	0	0	1	0
12	0	0	1	0
13	0	0	1	0
14	0	0	1	0

Questa codificazione viene chiamata *disgiuntiva*. Essa può essere proposta anche per i caratteri quantitativi. Se la variabile è discreta ed il numero r di modalità che essa può

assumere è piccolo, si procede come nel caso qualitativo. Se la variabile è continua oppure discreta ma presenta un grande numero di modalità, si suddivide l'intervallo in r classi a ciascuna delle quali si fa corrispondere una variabile indicatrice che risulta uguale a 1 se la generica h-esima unità presenta un valore compreso in tale classe e 0 altrimenti.

Se la codificazione disgiuntiva viene effettuata per tutti i p caratteri, la matrice dei dati originaria viene sostituita da una matrice $n \times P$ dove $P = \sum_{j=1}^{p} r_j$, dove r_j è il numero di modalità (o classi) della j-esima variabile.

1.5 Matrici a tre vie

Un'importante generalizzazione delle matrici di dati è costituita dalle cosidette *matrici a tre vie*. In termini generali, esse sono caratterizzate dal fatto che ciascun dato elementare che vi compare x_{hij} presenta tre indici che corrispondono ad una classificazione dello stesso in base a tre criteri:

$$\underset{n \times p \times q}{\mathbf{X}} = (x_{hij})$$

Esempi tipici sono i seguenti:

1. una matrice di dati del tipo "unità $\times$ variabili $\times$ occasioni" (ove il termine "occasioni" può indicare: diversi tempi, differenti situazioni sperimentali, luoghi diversi, etc.);

2. una successione di matrici di indici di prossimità del tipo "oggetti $\times$ oggetti" rilevate in differenti "'occasioni" (ad esempio, la similarità fra n prodotti di marche differenti, valutata dai consumatori in anni successivi, oppure in regioni diverse).

Un importante caso delle matrici a tre vie è costituito da quelle del tipo: "unità $\times$ variabili $\times$ tempi". In questo caso sulle stesse unità vengono rilevate le stesse variabili in più tempi successivi. La matrice di dati a tre vie può allora essere pensata come una successione temporale di nq matrici dei dati $\underset{\sim}{\mathbf{X}}_t = (x_{hit})$ del tipo consueto "unità $\times$ variabili". Alcuni esempi sono i seguenti:

1. la rilevazione in n aziende di p variabili desunte dai dati di bilancio in q anni successivi;

2. la misurazione in n pazienti di p sintomi, esprimibili quantitativamente, in q giorni;

3. i valori di un insieme di indicatori demografici, economici e sociali per ciascuna delle provincie italiane, rilevati dai Censimenti Istat nei vari anni.

I dati di questo tipo vengono chiamati *longitudinali* e la loro caratteristica essenziale è che le medesime unità statistiche sono "misurate" ripetutamente nel corso del tempo. Le analisi longitudinali estendono gli studi basati sulla matrice dei dati unità $\times$ variabili (spesso chiamati dati *sezionati* o di tipo *cross-section*), in cui per ogni unità statistica si dispone di un solo vettore di dati riferito ad un istante (o ad un intervallo temporale) prefissato. L'interesse delle matrici di dati a tre vie, in cui una delle dimensioni è il tempo, risiede nel fatto che esse permettono di misurare il cambiamento, con riguardo alle singole unità ed ai differnti fenomeni.

Un esempio economico di matrici a tre vie è fornito dalle azioni quotate in Borsa (le unità statistiche), dalle seguenti variabili: prezzo ufficiale, prezzo di riferimento (cioè calcolato sull'ultimo 10% delle contrattazioni giornaliere), numero di contratti effettuati, controvalore degli scambi, considerando più giorni successivi di Borsa aperta.

Se le unità statistiche delle matrici a tre vie formano un campione estratto da un certo universo, esse costituiscono un *panel*. Le indagini su dati provenienti da un *panel* costituiscono pertanto un caso particolare delle analisi longitudinali.

1.6 Trasformazioni lineari

Le trasformazioni lineari costituiscono uno dei principali strumenti dell'analisi di dati multidimensionali. Al fine dello studio di un fenomeno, in molti casi risulta più utile prendere in considerazione ed analizzare solo alcune combinazioni lineari delle variabili piuttosto che tutte le variabili originali, ciò poichè spesso si riduce la dimensione dei dati. Le trasformazioni lineari inoltre possono semplificare la struttura della matrice di varianze e covarianze, rendendo più semplice l'interpretazione dei dati.

Sia $\mathbf{a} = (a_1, \ldots, a_p) \in \mathbb{R}^p$ e consideriamo una combinazione lineare di $\mathbf{x}_h = (x_{h1}, \ldots, x_{hp}) \in \mathbb{R}^p$:

$$y_h = a_1 x_{h1} + \cdots + a_p x_{hp} = \mathbf{a}'\mathbf{x}_h \qquad\qquad h = 1, \ldots, n$$

dove $a_1, \ldots, a_p$ sono assegnati. In base alle (1.6) e (1.8) la media e la varianza dei y_h sono date rispettivamente da:

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{h=1}^{n} y_h = \frac{1}{n}\mathbf{a}'\sum_{h=1}^{n}\mathbf{x}_h \;=\; \mathbf{a}'\bar{\mathbf{x}} \\ s_y^2 &= \frac{1}{n}\sum_{h=1}^{n}(y_h - \bar{y})^2 \;=\; \frac{1}{n}\sum_{h=1}^{n}\mathbf{a}'(\mathbf{x}_h - \bar{\mathbf{x}})(\mathbf{x}_h - \bar{\mathbf{x}})'\mathbf{a} \;=\; \mathbf{a}'\mathbf{S}\mathbf{a}\ . \end{aligned}$$

In generale siamo interessati a trasformazioni lineari $\mathbb{R}^p \to \mathbb{R}^q$ del tipo:

$$\mathbf{y}_h \;=\; \mathbf{A}\mathbf{x}_h + \mathbf{b} \qquad\qquad h = 1, \ldots, n \tag{1.19}$$

che può essere scritta

$$\underset{\sim}{\mathbf{Y}} = \underset{\sim}{\mathbf{X}}\mathbf{A}' + \mathbf{1}\mathbf{b}' \tag{1.20}$$

dove $\underset{\sim}{\mathbf{Y}}$ è una matrice $n \times q$, $\mathbf{A}$ è una matrice $(q \times p)$ e $\mathbf{b}$ è un vettore q-dimensionale.

Il vettore delle medie e la matrice di varianze e covarianze dei $\mathbf{y}_h$ sono allora dati da:

$$\bar{\mathbf{y}} \;=\; \mathbf{A}\bar{\mathbf{x}} + \mathbf{b} \tag{1.21}$$

$$\begin{aligned} \mathbf{S}_y \;&=\; \frac{1}{n}\sum_{h=1}^{n}(\mathbf{y}_h - \bar{\mathbf{y}})(\mathbf{y}_h - \bar{\mathbf{y}})' = \frac{1}{n}\sum_{h=1}^{n}[\mathbf{A}\mathbf{x}_h + \mathbf{b} - (\mathbf{A}\bar{\mathbf{x}} + \mathbf{b})][\mathbf{A}\mathbf{x}_h + \mathbf{b} - (\mathbf{A}\bar{\mathbf{x}} + \mathbf{b})]' \\ &=\; \mathbf{A}\mathbf{S}\mathbf{A}'\ . \end{aligned} \tag{1.22}$$

In particolare se $q = p$ e $\mathbf{A}$ è non singolare[11] , allora

$$\mathbf{S} \;=\; \mathbf{A}^{-1}\mathbf{S}_y(\mathbf{A}')^{-1} \; . \tag{1.23}$$

Esistono numerose trasformazioni lineari di interesse in statistica, qui di seguito ne vedremo alcuni. Per semplicità tutte le trasformazioni lineari sono centrate intorno allo $\mathbf{0}$.

1.6.1 La trasformazione di scala

Consideriamo la trasformazione $\mathbb{R}^p \to \mathbb{R}^p$ data da:

$$z_{hi} = \frac{x_{hi} - \bar{x}_i}{s_i} \qquad \text{o, in termini vettoriali} \quad \mathbf{y}_h \;=\; \mathbf{D}^{-1}(\mathbf{x}_h - \bar{\mathbf{x}}) \tag{1.24}$$

per $h = 1, \ldots, n$, $i = 1, \ldots, p$ con $\mathbf{D}$ =diag(s_i). La (1.24) opera un cambiamento di scala in modo tale che ciascuna variabile abbia varianza unitaria e pertanto si elimina l'arbitrarietà nella scelta della scala. Le quantità z_{hi} vengono chiamate *scarti standardizzati*. Si dimostra che, in forma matriciale, la (1.24) si può scrivere:

$$\underset{\sim}{\mathbf{Z}} = \mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} \; .$$

dove $\underset{\sim}{\mathbf{Z}}' = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. Si ha infatti dalla (1.24)

$$\mathbf{z}_h = \mathbf{D}^{-1}\mathbf{x}_h - \mathbf{D}^{-1}\bar{\mathbf{x}}$$

da cui, in base alla (1.20) e tenendo conto che $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$, si ha:

$$\begin{aligned}\underset{\sim}{\mathbf{Z}} &= \underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} + \mathbf{1}(\mathbf{D}^{-1}\bar{\mathbf{x}})' = \underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} + \mathbf{1}(\mathbf{D}^{-1}\frac{1}{n}\mathbf{X}'\mathbf{1})' = \underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} + \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\mathbf{D}^{-1} \\ &= (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} = \mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} \; .\end{aligned}$$

Da un'applicazione di tale risultato – poichè $\mathbf{H1} = \mathbf{0}$ – segue:

$$\begin{aligned}\bar{\mathbf{z}} &= \frac{1}{n}\underset{\sim}{\mathbf{Z}}'\mathbf{1} = \frac{1}{n}\mathbf{D}^{-1}\underset{\sim}{\mathbf{X}}'\mathbf{H}\mathbf{1} = \mathbf{0} \\ \mathbf{S}_z &= \frac{1}{n}\underset{\sim}{\mathbf{Z}}'\mathbf{H}\underset{\sim}{\mathbf{Y}} = \frac{1}{n}\mathbf{D}^{-1}\underset{\sim}{\mathbf{X}}'\mathbf{H}'\mathbf{H}\mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} = \frac{1}{n}\mathbf{D}^{-1}\underset{\sim}{\mathbf{X}}'\mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{S}_x\mathbf{D}^{-1} = \mathbf{R} \, ,\end{aligned}$$

[11] **Definizione.** Una matrice quadrata $\mathbf{A}$ si dice *non singolare* se $|\mathbf{A}| \neq 0$; altrimenti se $|\mathbf{A}| = 0$ la matrice $\mathbf{A}$ si dice singolare.

Alcune proprietà delle matrici non singolari. Sia $\mathbf{A}$ una matrice quadrata non singolare e $\alpha \in \mathbb{R}$. Allora si ha:

(i) $\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}(A_{ij})'$.

(ii) $(\alpha\mathbf{A})^{-1} = \frac{1}{\alpha}\mathbf{A}^{-1}$.

(iii) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

(iv) L'unica soluzione di $\mathbf{Ax} = \mathbf{b}$ `e $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

(v) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

(vi) $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$.

essendo **H** una matrice idempotente.

Ad esempio, con riferimento al file dati bordeaux, riportiamo i valori trasformati in accordo alla (1.24) per le prime dieci osservazioni:

Annata	Z_1	Z_2	Z_3	Z_4	Qualità
24	-0.6753	-0.3813	-0.8941	0.0062	+
25	-1.1313	-1.5641	-0.7928	-0.2492	-
26	-0.0268	-0.9247	0.0179	0.3616	+
27	-0.5256	-2.2274	-1.5022	1.1834	–
28	0.6145	0.0743	1.7406	-0.7378	++
29	0.7713	1.0973	1.6393	-1.5041	+++
30	-0.5613	-2.2594	-0.5901	0.6281	–
31	-1.3166	-0.5091	-0.6915	1.4166	–
32	-0.8606	-1.1645	-0.4888	3.5155	—
33	1.1347	0.4899	1.0312	0.7392	+

1.6.2 La trasformazione di Mahalanobis

Se $\mathbf{S} > 0$ allora $\mathbf{S}^{-1}$ ha un'unica radice quadrata $\mathbf{S}^{-1/2}$. La trasformazione di Mahalanobis è $\mathbb{R}^p \to \mathbb{R}^p$ è definita da:

$$\mathbf{y}_r \;=\; \mathbf{S}^{-1/2}(\mathbf{x}_r - \bar{\mathbf{x}}) \qquad\qquad h = 1, \ldots, n\ . \tag{1.25}$$

Analogamente a quanto fatto in precedenza, in forma matriciale, la (1.25) si può scrivere:

$$\underset{\sim}{\mathbf{Y}} \;=\; \mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{S}^{-1/2}\ , \tag{1.26}$$

dove $\underset{\sim}{\mathbf{Y}}' = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$. Segue poi, con ragionamenti analoghi a quelli visti in precedenza:

$$\bar{\mathbf{y}} = \frac{1}{n}\underset{\sim}{\mathbf{Y}}'\mathbf{1} = \frac{1}{n}\mathbf{S}^{-1/2}\underset{\sim}{\mathbf{X}}\mathbf{H}'\mathbf{1} = \mathbf{0}$$

$$\mathbf{S}_y = \frac{1}{n}\underset{\sim}{\mathbf{Y}}'\mathbf{H}\underset{\sim}{\mathbf{Y}} = \frac{1}{n}\mathbf{S}_x^{-1/2}\underset{\sim}{\mathbf{X}}'\mathbf{H}'\mathbf{H}\mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{S}_x^{-1/2} = \frac{1}{n}\mathbf{S}_x^{-1/2}\underset{\sim}{\mathbf{X}}'\mathbf{H}\underset{\sim}{\mathbf{X}}\mathbf{S}_x^{-1/2} = \mathbf{S}_x^{-1/2}\mathbf{S}_x\mathbf{S}^{-1/2} = \mathbf{I}\ .$$

Pertanto tale trasformazione elimina la correlazione fra le variabili standardizza la varianza di ciascuna variabile.

1.6.3 Trasformazione in componenti principali

Poichè la matrice di varianze e covarianze **S** è simmetrica, allora in base al teorema di decomposizione spettrale[12]

può essere scritta nella forma:

$$\mathbf{S} \;=\; \mathbf{G}\mathbf{L}\mathbf{G}' \tag{1.31}$$

[12] **Teorema della decomposizione spettrale.** Sia **A** una matrice quadrata simmetrica di ordine $n \times n$. Allora la matrice **A** pu`o sempre essere scritta come

$$\mathbf{A} \;=\; \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'$$

L è la matrice diagonale degli autovalori di **S**, con $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$ e **G** è una matrice ortogonalele avente per colonne i corrispondenti autovettori di **S**. La trasformazione in componenti principali $\mathbb{R}^p \rightarrow \mathbb{R}^p$ è definita dalla rotazione

$$\mathbf{w}_r \;=\; \mathbf{G}'(\mathbf{x}_r - \bar{\mathbf{x}}) \qquad\qquad h = 1, \ldots, n \; . \tag{1.32}$$

Poichè $\mathbf{S}_w = \mathbf{G}'\mathbf{S}\mathbf{G} = \mathbf{L}$ è diagonale, le colonne di **W**, chiamate *componenti principali*, sono delle combinazioni lineari non correlate delle variabili. In pratica si spera di riassumere la maggior parte della variabilità nei dati usando solo le componenti principali con la maggiore varianza, così da ridurre la dimensione

Poichè le componenti principali sono non correlate con varianze $l_1, \ldots, l_p$ sembra naturale definire una dispersione totale dei dati mediante una funzione crescente di $l_1, \ldots, l_p$ come ad esempio $\prod l_i$ oppure $\sum l_i$. Poichè $|\mathbf{S}| = |\mathbf{L}| = \prod l_i$ e tr**S** =tr**L** $= \sum_i l_i$, la trasformazione in componenti principali fornisce una motivazione per le misure di dispersione introdotte in precedenza.

1.7 Aspetti geometrici

Ci sono due diverse modalità in base alle quali possiamo studiare una matrice di dati.

Una prima analisi concerne lo studio delle colonne della matrice $\underset{\sim}{\mathbf{X}}$, cioè delle variabili. Ciò conduce ad un insieme di metodologie dette *tecniche-R* in quanto un ruolo importante viene svolto dalla matrice di correlazione **R**. Esempi sono l'analisi in componenti principali, l'analisi fattoriale e l'analisi delle correlazioni canoniche.

La matrice dei dati $\underset{\sim}{\mathbf{X}}$ può essere anche studiata confrontando le righe, cioè i diversi elementi o oggetti. Ciò porta a tecniche quali l'analisi discriminante, l'analisi dei gruppi (cluster analysis) e lo scaling multidimensionale.

Questi due diversi approcci corrispondono a due diversi punti di vista geometrici di rappresentare la matrice di dati $(n \times p)$. Nel primo caso le colonne possono essere viste

---

dove $\mathbf{\Lambda}$ `e la matrice diagonale degli autovalori di **A** e $\mathbf{\Gamma}$ `e una matrice ortogonale le cui colonne sono i corrispondenti autovettori standardizzati di **A**.

Corollario. Sia **A** una matrice simmetrica non singolare. Allora per ogni n si ha:

$$\mathbf{\Lambda}^n \;=\; \text{diag}(\lambda_i^n) \qquad\qquad \text{e} \qquad\qquad \mathbf{A}^n \;=\; \mathbf{\Gamma}\mathbf{\Lambda}^n\mathbf{\Gamma}' \; . \tag{1.27}$$

Se tutti gli autovalori di **A** sono positivi, allora si possono defi nire le potenze razionali:

$$\mathbf{A}^{r/s} \;=\; \mathbf{\Gamma}\mathbf{\Lambda}^{r/s}\mathbf{\Gamma}' \qquad\qquad \text{dove} \qquad\qquad \mathbf{\Lambda}^{r/s} \;=\; \text{diag}(\lambda_i^{r/s}) \; , \tag{1.28}$$

per ogni intero $s > 0$ e r. Se qualche autovalore di **A** `e uguale a zero, allora le (1.27) e (1.28) valgono se gli esponenti sono non negativi.

Importanti casi particolari della (1.28) sono

$$\mathbf{A}^{1/2} \;=\; \mathbf{\Gamma}\mathbf{\Lambda}^{1/2}\mathbf{\Gamma}' \qquad , \qquad \mathbf{\Lambda}^{1/2} \;=\; \text{diag}(\lambda_i^{1/2}) \; , \tag{1.29}$$

con $\lambda_i \geq 0$ per ogni i, e

$$\mathbf{A}^{-1/2} \;=\; \mathbf{\Gamma}\mathbf{\Lambda}^{-1/2}\mathbf{\Gamma}' \qquad , \qquad \mathbf{\Lambda}^{-1/2} \;=\; \text{diag}(\lambda_i^{-1/2}) \; . \tag{1.30}$$

La decomposizione (1.29) `e chiamata *decomposizione della radice quadrata simmetrica* di **A**.